Research Paper

Node2vec Representation for Clustering Journals and as A Possible Measure of Diversity

Zhesi Shen[1], Fuyou Chen[1], Liying Yang[1], Jinshan Wu[2†]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190, P.R.China
[2]School of Systems Science, Beijing Normal University, Beijing, 100875, P.R.China

Citation: Zhesi Shen, Fuyou Chen, Liying Yang, Jinshan Wu (2019). Node2vec representation for clustering journals and as a possible measure of diversity. Journal of Data and Information Science, 4(2), 79–92

DOI: 10.2478/jdis-2019-0010

Received: Apr. 3, 2019
Revised: Apr. 26, 2019
Accepted: May 9, 2019

Abstract

Purpose: To investigate the effectiveness of using node2vec on journal citation networks to represent journals as vectors for tasks such as clustering, science mapping, and journal diversity measure.

Design/methodology/approach: Node2vec is used in a journal citation network to generate journal vector representations.

Findings: 1. Journals are clustered based on the node2vec trained vectors to form a science map. 2. The norm of the vector can be seen as an indicator of the diversity of journals. 3. Using node2vec trained journal vectors to determine the Rao-Stirling diversity measure leads to a better measure of diversity than that of direct citation vectors.

Research limitations: All analyses use citation data and only focus on the journal level.

Practical implications: Node2vec trained journal vectors embed rich information about journals, can be used to form a science map and may generate better values of journal diversity measures.

Originality/value: The effectiveness of node2vec in scientometric analysis is tested. Possible indicators for journal diversity measure are presented.

Keywords Science mapping; Diversity; Graph embedding; Vector norm

1 Introduction

Mapping the structure of science into hierarchical clusters/communities provides the basis of many further scientometrical investigations. For example, citation field normalization depends on the clusters (Waltman, 2016). Evolution of science disciplines and the interrelation among them can be better captured at the level of clusters rather than at the level of individual journals or individual papers (Shen

[†] Corresponding author: Jinshan Wu (E-mail: jinshanw@bnu.edu.cn).

et al., 2016). Furthermore, even a simple visualization of the map of science can be informative. Such clustering can be done in citation networks of either journals or papers via algorithms (Leydesdorff, 2006; Leydesdorff et al., 2017; Waltman & Van Eck, 2012), or even manually (Glänzel & Schubert, 2003). The cutting edge algorithm and studies of paper-level clustering is the Smart Local Moving (SLM) algorithm and related analysis (Boyack & Klavans, 2014; Colavizza et al., 2018; Haunschild et al., 2018; Klavans & Boyack, 2017; Sjogarde & Ahlgren, 2018).

The first research question that we are tackling here is to find another clustering algorithm of journals/papers. Or more specifically mainly in this work, we focus on a sub-component of a clustering algorithm, which is the similarity measure of journals/papers. One thing we notice in earlier clustering algorithms is that it is better to consider both text-based similarity and citation-based similarity (Boyack et al., 2017; Glänzel & Thijs, 2011; Janssens et al., 2008). Among the clustering algorithms developed for networks, node2vec/word2vec naturally takes both into consideration (Grover & Leskovec, 2016; Mikolov et al., 2013). Starting from a language corpus, a machine learning algorithm word2vec generates a vector representation of each word based on the assumption that words appearing close to each other are more similar in their meaning. Node2vec first converts a network into a corpus of nodes of the network and then applies word2vec to generate a vector for each node of the network. The vectors can then be used to calculate the similarity between each pair of the words or of the nodes. Thus, a combination of node2vec and word2vec, which are in fact just one algorithm in their cores, is a natural hybrid clustering method of both text-based and citation-based similarity of papers or journals. Once we have a similarity measure, in principle, we may apply various algorithms for clustering. In this work, however, since we focus more on similarity measure, we will use simply the hierarchical clustering (Pedregosa et al., 2011) as the algorithm of clustering.

Such a paper-level clustering, which is already an ongoing study in our group, requires citation data, text data of all papers, and tedious work on analysis and also on examining/validating the results. Thus, in this work, we would like to report our results on the node2vec/word2vec on journals first. In fact, at the level of journals, we only consider the citation network of journals, but not any texts from the journals, since it is a tricky question to define representative texts of journals. When the same method is applied to paper-level clustering, texts of a paper can be its title and abstract or even its full text.

As we will show later, the generated map of science has a reasonably clear community structure and the calculated communities of most of the journals agree well with several of the existing classifications of journals. Once we have the map of science, we notice that when we visually examine the map, overall, the multidisciplinary journals are at the boundary area of several disciplines and they are closer to the centers than to the edges. This inspires us to look at the vectors as a potential source for measuring the diversity/multidisciplinarity of journals.

Therefore, the second research question we investigate in this work is the plausibility of using the vector of each journal obtained via the node2vec for the diversity measure of the journal. This first relates to the question of how the journal vector itself, specifically the vector norm, correlates with journal diversity. Second, this plausibility might also depend on which similarity measure is used in the calculation of diversity. Often the definition of diversity of either journals or papers relies on a measure of similarity, see, for example, the Rao-Stirling definition in Eq. (2). This similarity can be defined based on co-cited or co-citing papers/journals. For example (Leydesdorff et al., 2018), for journal *i* and journal *j*, their similarity can be calculated based on their vectors of direct citations $v_i^c = \left[c_1^i, c_2^i, \cdots, c_N^i\right]^T$, where c_m^i is the number of citations from journal *i* to journal *m*.

It has been pointed out that the similarity measure plays an important role in diversity measure (Zhang et al., 2016). Since we now have another vector representation, and in some sense, the vector has the property that similar journals have similar vectors, we replace the citation vector $v_i^{c,(1)}$ with the node2vec trained vector in measuring the diversity of journals. We show that this substitution significantly improves the validity of Rao-Stirling diversity.

2 Method and Data

The journal citation network is extracted from the Journal Citation Reports 2017 (JCR2017, 2018). Links with citation counts less than 6 are removed. The resulted journal citation network has about 11,000 nodes and 950,000 links.

Node2vec is a representational learning framework of graphs, which can generate continuous vector representations for the nodes based on the network structure (Grover & Leskovec, 2016). The core algorithm in node2vec is word2vec (Mikolov et al., 2013). Here we use node2vec to learn 32-dimensional vectors v^n for each journal based on the journal citation network. In fact, we also tested 64 and 128-dimensional vector representation and we found similar map of science and similar clusters of journals. Please see the caption of Fig. 2(b) to further details. The underlying mechanism of node2vec is to produce vectors of nodes so that the nodes having more citation links together should have more similar vector representations. Thus, later those nodes will have larger similarity as the dot product of similar vectors is larger then the dot product of very different vectors. This mechanism fits fairly well to the assumption of other clustering algorithms, for example Waltman and Van Eck (2012). Thus, in a sense, it is not surprising if vectors generated from node2vec can be used in clustering journals/papers. Where then this good character of node2vec comes from? This roots in the word2vec, which is the basis of node2vec. The essential assumption of word2vec, is that the words that appearing often close to each other have related meaning and thus a good algorithm of vector representation of words should also lead to similar vectors for such words. In the case of the word, then each of the dimensions of the vector space represents a sense of meaning and the generated vector presentation of a word means in such a vector space of meaning, where the word locates.

We basically extract the journal citation data and run it through the node2vec algorithm to get vectors of predefined dimensions, and then use the vectors to calculate journal pairwise similarity for clustering, and also feed the vectors into t-SNE (Maaten & Hinton, 2008) to get a 2-d presentation for science mapping. Here we set the dimension parameter as 32 considering both performance and computational cost and comparisons against other dimension parameters are shown in Fig. 2. All code is available at https://github.com/challenge19/Journal_node2vec.

Rao-Stirling diversity indicator is one of the widely used diversity measurements, which takes variety, balance, and disparity into consideration. The Rao-Stirling diversity is calculated as follows,

$$\begin{aligned} D &= \sum_{i \neq j} p_i p_j \left(1 - S_{ij}\right), \\ S_{ij} &= \max\left[0, \cos(\vec{v}_i, \vec{v}_j)\right]. \end{aligned} \tag{1}$$

where the vector used to calculate similarity can be either the direct citation vector v_i^c or the node2vec trained vector v_i^n of each journal,

$$\begin{aligned} v_i^c &= \left[c_1^i, c_2^i, \cdots, c_L^i\right]^T, \\ v_i^n &= \left[v_1^i, v_2^i, \cdots, v_N^i\right]^T. \end{aligned} \tag{2}$$

Here c_j^i is the citation counts between journal i and journal j, L is the total number of journals considered and N is the number of dimensions of the vector space representing all journals resulted from node2vec.

3 Results

3.1 Clustering and Science mapping

To test the effectiveness of the trained vector from node2vec, we first make a science map of the journals according to their vectors by t-SNE, which reduces the vector dimension to 2. Presentation of our science map is shown in Fig. 1, in which each colored dot represents a journal with its color encoding its category used in

Essential Science Indicators (ESI). ESI journal category is a well known and commonly used journal classification system. Since the ESI journal category is not directly generated by our method, we think it is more objective to use it to evaluate the reasonableness of our science map, especially the closeness of journals within same categories and the relative positions of categories.

First, we notice that our science map has a clear community structure and has a high agreement with the ESI journal category: journals with the same colors (same category) are more or less concentrated near their own region in the 2-dimensional t-SNE plot. This implies that the similarity of vectors among journals in the same category is much larger than journals across categories. We also noticed high overlaps between BIOLOGY & BIOCHEMISTRY journals and MOLECULAR BIOLOGY & GENETICS journals, implying that these journals are densely connected and may be merged into one category.

Second, we also find that in general, the structure of this science map is quite similar to earlier maps (Klavans & Boyack, 2009), with similar positions and relationships among disciplines, except for Geoscience. For example, Economics is close to other Social Sciences, Math, and Computer Science, and Material Science is between Physics, Engineering, and Chemistry, while Biology and Medicine are very close to each other, and Environmental Science is close to Chemistry, Biology, and Geoscience. Geoscience in our map is located peripherally, while in several other maps it is usually near the center. Intuitively, Geoscience is a discipline that makes use of many other disciplines, thus it is understandable that it is near the center. Yet, placing Geoscience near Environmental Science, Chemistry and Agriculture also makes sense. Therefore, we conclude that even for the position of Geoscience both our map and other maps capture part of the real connections among scientific disciplines.

Journals indexed as Multidisciplinary are highlighted on the map as red dots with black borders. Instead of being at the center, which means that there is no bias towards any disciplines, we see that these so-called multidisciplinary journals almost scatter all across the map with a quite several of them located at the boundary area of medicine, microbiology and related life sciences. This means that many of those multidisciplinary journals have their own focuses. For example, *Nature*, *Science*, *PNAS* and *Nat. Commun.* locate closely in the biology cluster and this implies that they have similar discipline formation pattern and mainly focus on biology. *PLoS One* locates at the border of Medicine and Biology because its majority publications are biomedicine oriented. *Nat. Sci. Rev.*, hosted by Chinese Academy of Sciences and mainly review cutting-edge developments across science and technology in China, locates at the physical science part as China is relatively strong in Chemistry, Material Science and Physics. Although these journals all are indexed as multidisciplinary, their published papers present heterogeneous orientations. This also implies that, when ranking journals, simply putting them together as a set of multidisciplinary journals might be inappropriate.

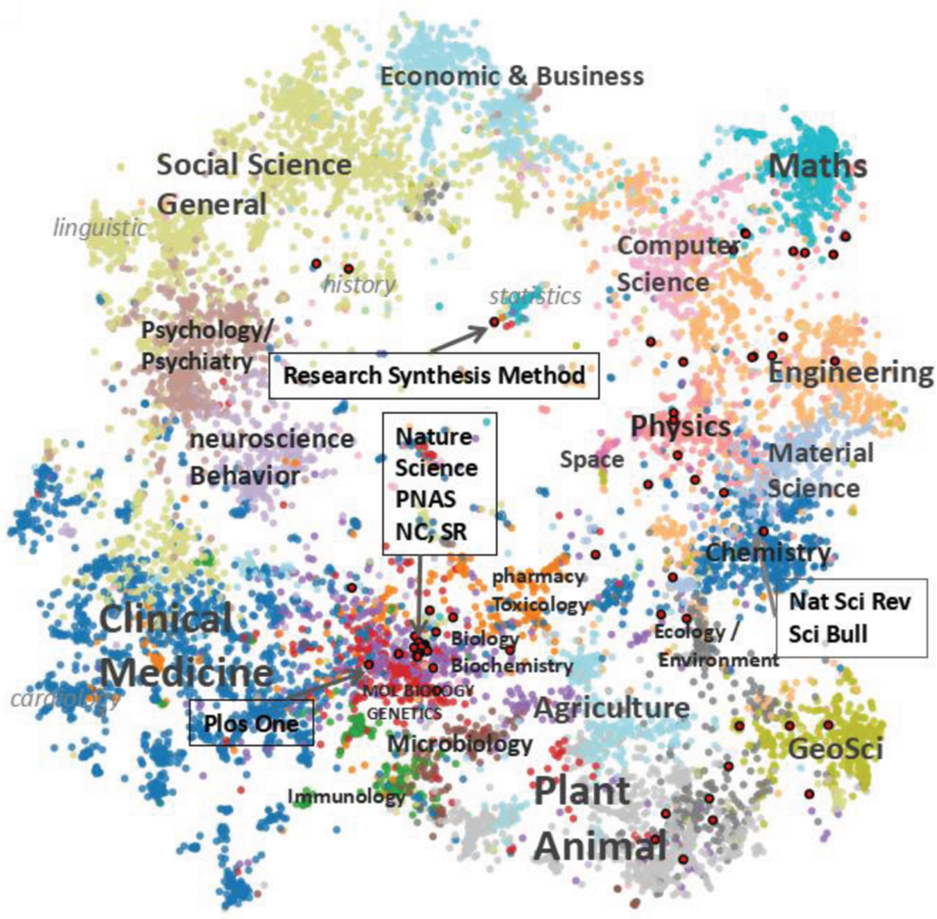

Figure 1. Map of scientific journals. Colors of dots mean the corresponding ESI categories of journals. Dots in red with black border are journals indexed as multidisciplinary, of which we list only a few on the map.

We further cluster the journals at various levels of aggregations with hierarchical clustering (Pedregosa et al., 2011). Let us refer to our system of hierarchical clusters of journals as the Vec clusters (Vec for short). We then compare Vec with several commonly used journal classification systems, e.g., JCR-WoS subject categories[①] (JCR), automated journal classification using VOSviewer (Leydesdorff et al., 2017)

[①] http://mjl.clarivate.com/scope/scope_scie/

(VOS), ESI disciplines[②] (ESI) and LCAS-journal-ranking journal categories[③] (LCAS). Adjust Mutual Information (AMI) (Vinh et al., 2010) is used to measure the agreement between two classification systems. AMI is a similarity measure of two classifications, meaning high values of AMI the two classifications are more similar.

Since these four classification systems of journals are aggregated at different levels, e.g. ESI with 22 fields and JCR with about 200 disciplines, we generate Vec across different levels and compare them with these classification systems. Note that each of these pre-existing classifications has fixed number of clusters while Vec is generated at various granularity. As shown in Fig. 2(a), we can see a moderate agreement between Vec and these classification systems. The AMI shows a peak when the two compared classifications are aggregated at similar levels since at this common size of clusters the two classifications are often more similar than the cases when their sizes are different. We see that, overall, the agreement between Vec and VoS is higher than others. This is due to the fact that they both use journal citation data to generate the classification systems.

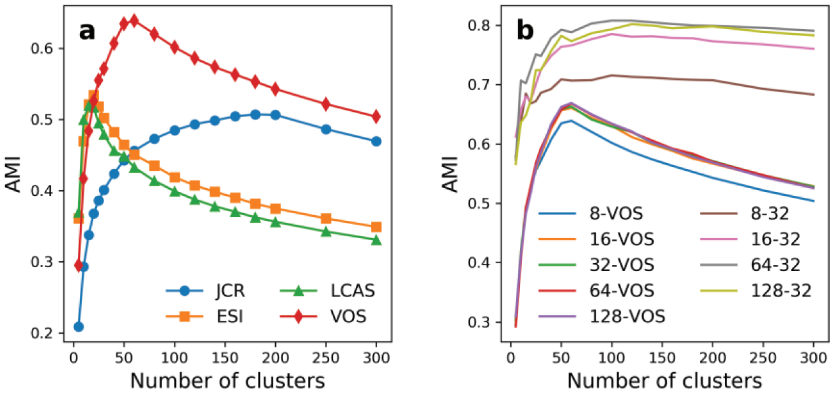

Figure 2. (a) Our vector-based clustering of journals compared with several existing journal classification systems JCR, VOS, ESI and LCAS. (b) We also compare resulted clusters using various dimensions of the node2vec vectors and we find there is not much differences among $d = 32$, $d = 64$ and $d = 128$: the similarity of 64–32 and 128–32 are much higher than that of 8–32 while 16–32 is somewhere in between. Also when Vec clusters with $d = 8, 16, 32, 64, 128$ are compared against VOS, we find as long as $d > 16$, increasing d does not make a big difference. Considering both performance and computational cost, we only report the results of $d = 32$.

Both the science map and this comparison of the maps show that the generated node2vec vectors of journals do capture part of inherent characters of each journal

[②] https://help.incites.clarivate.com/incitesLiveESI/ESIGroup/overviewESI/esiJournalsList.html
[③] http://www.fenqubiao.com/

such that the resulted clusters are quite reasonable. Based on this observation, we want to further examine these vectors.

3.2 Journal analogy test like the "King - Man + Woman = Queen" of word analogy test

One advantage of word2vec-like algorithms is its vector will embed the semantic relationships among words. The most famous example is the "King - Man + Woman = Queen" relation. Here we also test such a phenomenon on the trained vector of journals. Table 1 shows the result of such a test. Here we fixed "King = *PLoS Comput. Biol.*", "Man = *Nat. Cell Biol.*", and test three journals from different disciplines as "Woman". *PLoS Comput. Biol.* is more focusing on the computational and mathematical parts of biology than *Nat. Cell Biol.*. With such a difference between them, after *Phys. Rev. Lett.* added, it is expected that the "Queen"-like journals will be more towards computational and mathematical physics. Indeed, we find journals more or less concentrated on computational/mathematical physics: *J. Stat. Mech. Theory Exp.*, *Phys. Rev. E* and *Eur. Phys. J. B.* clearly have this character. Similarly, when "Woman = *Genome Biol.*", the resulted journals mainly focus on bioinformatics, the computational part of genome study; and when "Woman = *J. Neurosci.*", the most similar journals are computational neuroscience journals. In all the above cases, the mathematical and computational character of the "King = *PLoS Comput. Biol.*" is more or less kept in the Queen journals. This also partially validates that the trained vectors encode to some degree the intrinsic characteristics of journals.

Table 1. The "King - Man + Woman = Queen" test on the node2vec trained vectors of journals. Top-5 "Queen"-like journals are presented.

Example	Test 1	Test 2	Test 3
King		*PLoS Comput. Biol.*	
Man		*Nat. Cell Biol.*	
Woman	*Phys. Rev. Lett.*	*Genome Biol.*	*J. Neurosci.*
Queen	*J. Stat. Mech. Theory Exp*	*Bioinformatics*	*NeuroImage*
	Phys. Rev. E	*BMC Bioinformatics*	*Biol. Cybern.*
	Fluctuation Noise Lett.	*J. Comput. Biol.*	*Front. Comput. Neurosci.*
	EPL	*BioData Min.*	*Cereb. Cortex*
	Eur. Phys. J. B	*J. Bioinform. Comput. Biol.*	*J. Comput. Neurosci.*

3.3 Vector norm and diversity of journals

Roughly speaking each dimension of the node2vec vectors likely represents a specific subject at a certain level depending on the chosen parameter of dimensions of the vector space. Thus, a more specified journal might cover only a few such directions while a journal with a broader scope might be seen as a summation of several of those directions. Then the summation of a few vectors may still point to some specific directions with larger length, while the summation of many vectors, each pointing to different directions, will often be short and not pointing to some specific directions. This is very much like the in central limit theorem that the more random variables in the summation the closer to 0 is the average. In fact, in the vectors of words generated from word2vec, vector norm of words with multiple meanings and different contexts is often shorter than that of the words with very specific meaning and context (Schakel & Wilson, 2015). Based on this intuition, we want to test the plausibility of using vector norm as an indicator of the diversity of journals.

To do so, we first compare the values of vector norms of the commonly known multidisciplinary journals against that of other journals. In Fig. 3, overall, we can see that those indexed as multidisciplinary journals (orange dots) mainly are near the lower part of the scatter plot, such as *PLoS One*, *Sci. Rep.*, *Nat Commun*, *Nature* and so on, meaning that often they do have smaller norms. More specific journals such as *Fem. Stud.*, *J. Topol.*, *Brit. J. Relig. Educ.* and so on, do have larger norms. This in fact is sufficient to illustrate our point that multidisciplinary journals overall have smaller vector norms than more specific journals.

However, in Fig. 3 we also use another journal indicator: node centrality of journals. Node centrality is measured by the occurrence frequency in the random walk series generated for node2vec, which can be treated as a network centrality measure and is closely related to node degree or the total received citations of journals in the citation network. Thus, node centrality of a journal is similar to the total impact of the journal. We find that even for journals with similar node centrality, the one with the smaller vector norm is more likely a multidisciplinary journal. For example, on the right part, the three journals – *Nat. Commun.*, *Phys. Rev. Lett.* and *Astro. J.* – have similar node centrality values, the vector norm of *Nat. Commun.* is smaller than that of *Astro. J.* with *Phys. Rev. Lett.* appearing in the middle, as the topic broadness of these journals is *Nat. Commun.* > *Phys. Rev. Lett.* > *Astro. J.*. Similar situation also holds for journals with medium node centrality, e.g., *Sci. Data*, *Res. Synth. Methods* and *J. Topol.*. Interestingly, we see again that *Endeavour* and *Int. J. Bifurcation Chaos*, although indexed as multidisciplinary journals, have relatively large vector norms compared with journals with similar node centrality.

Thus, it seems that vector norm indeed encodes diversity character to some degree. Next, we would like to develop a practical indicator of journal diversity using the vector representation of journals. However, finding the best way to measure journal diversity is still an open question. Often, for a measure of journal diversity, citing vector of a journal, which means the given journal cites how many times each of other journals or cited vector of a journal, which means how many

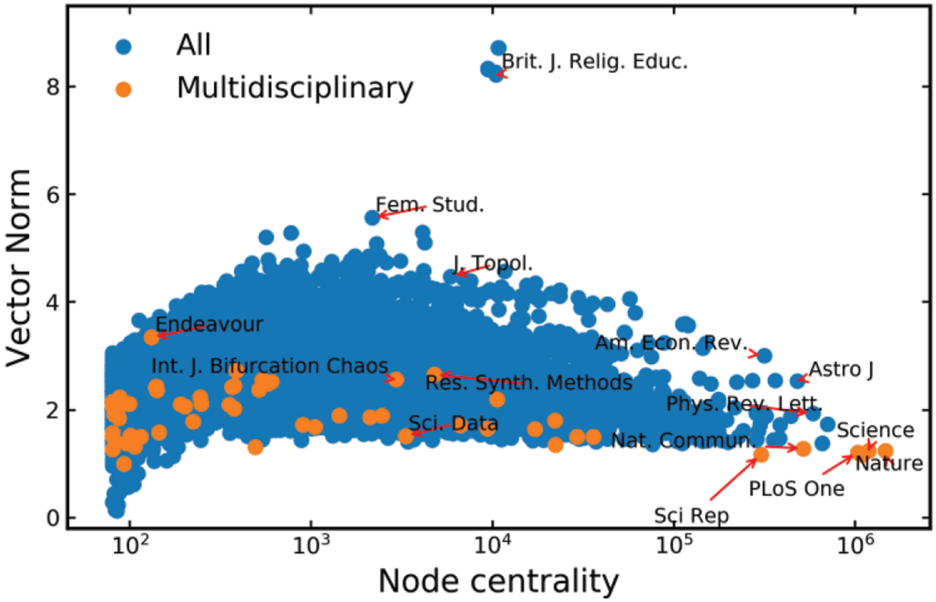

Figure 3. Scatter plot of vector norms versus node centrality. Node centrality is measured as the node occurrence frequency in the random walk series generated for Node2Vec. Orange dots represent journals indexed in Multidisciplinary Science.

times this journal is cited by each of other journals, or both are needed. Since our goal in this work is not really to develop such an indicator but rather examining the usefulness of node2vec vectors of journals, we take simply one of the widely used indicators of diversity, the Rao-Stirling diversity (Rao, 1982; Stirling, 2007) in Eq. (2) and replaced the citation vector v^c with our node2vec vectors v^n. We then check whether or not such a replacement improves diversity values.

Following the earlier settings (Leydesdorff et al., 2018), we calculate the citing-side and cited-side Rao-Stirling diversity of all journals in our set, as shown in Fig. 4. In Fig. 4(a), node2vec vectors are used to calculate journal similarity and in Fig. 4(b), we use the citation vectors. In Fig. 4, the x-axis "Citing diversity" means the diversity of the citing vector defined above and the y-axis "Cited diversity" means the diversity of the cited vector defined above.

Overall, we see that in Fig. 4(a) when node2vec vectors are used, many of the commonly known multidisciplinary journals are at the diagonal top region of the plot and separated from other journals, especially those clearly multidisciplinary ones, such as *Nature*, *Science*, *PLoS One*, and *Sci. Rep.*. However, in Fig. 4(b) when we use citation vectors, almost all multidisciplinary journals are in the cloud of other journals and there is no clear separation at all. Again, we are not highlighting any definition of diversity measure in this work. Instead, we are trying to point out that replacing the citation vector with the node2vec vectors might help to find a better indicator of diversity, even when the definition of diversity, like the Rao-Stirling diversity, is still the same. We just replaced the measure of similarity/disparity.

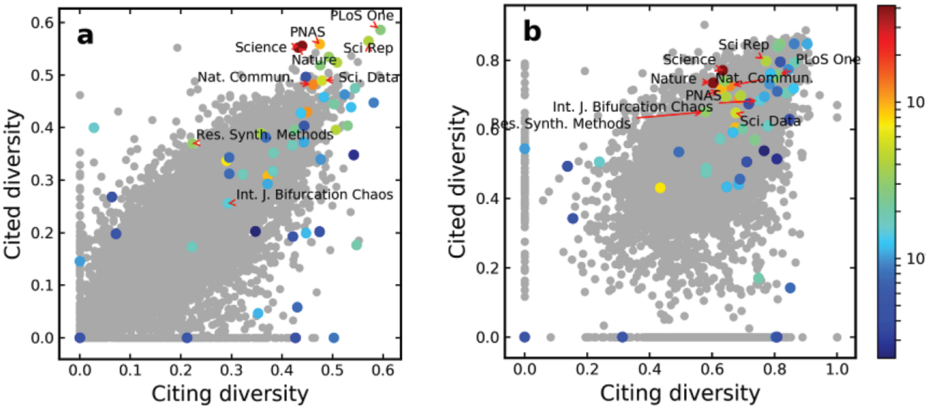

Figure 4. Diversity of journals calculated using similarity measured by (a) vector v^n learned from node2vec and (b) vector v^c. Journals indexed as Multidisciplinary are colored according to their JIFs with blue implying low JIF and red implying high JIF as shown in the right legend. The citing diversity of journal i is measured based on its referenced journals, and cited diversity is measured based on the journals citing it.

4 Conclusion and Discussion

From this test of the node2vec algorithm on the citation network of journals, we see that the node2vec-based science map of journals has relatively clear separations, the closeness among the clusters represents true relationships among corresponding disciplines, and the overall structure of the map agrees with other commonly used maps of sciences. This means that the journal vectors generated from node2vec capture essential characteristics of these journals. This is also illustrated in Table 1.

We also found that it is possible to use the norm of generated vectors for diversity measurement. It was at first surprising to us that vectors of multidisciplinary journals have smaller norms. However, similar things happen for the word2vec vector representation of words. Words whose meanings are very specialized have a big norm, while words whose meanings are rather common have a smaller norm (Schakel & Wilson, 2015). We suspect that the same thing happens here so that the more specialized journals have larger norms. Based on this observation, we tested the applicability of the vector representation to diversity measurement. Using the Rao-Stirling diversity measure as an example, we show that replacing the often-used citation vectors with our node2vec vectors indeed improves the agreement between the measured values of diversity and the commonly known multidisciplinary journals. When node2vec vectors are used, several multidisciplinary journals become outliers since their measured diversity values are larger than other journals, while almost all multidisciplinary journals are not distinguishable from other journals when citation vectors are used.

Based on the above observations, we conclude that node2vec/word2vec has potential in clustering and characterizing nodes in citation networks. Simply due to the complexity of analysis, in this work, we decide to present results on journal-level science mapping first. Extending the current study to a paper-level network, which will naturally integrate citation networks and text contents of papers, will be the topic of future investigations. Such an extension might contribute to both the method of paper clustering and paper diversity measurement. Also there we do not study which diversity measure is the better.

To illustrate the main concepts and contribution, we prepared a graphic summary of this work in Fig. 5. The concepts and connections among concepts are color coded. The black ones are from references (Grover & Leskovec, 2016; Mikolov et al., 2013). The current work implements only the red ones and the green ones about paper-level clustering is a topic of future investigation. Preliminary results of this work were reported at the 2nd International Conference on Data-driven Knowledge Discovery in 2018.

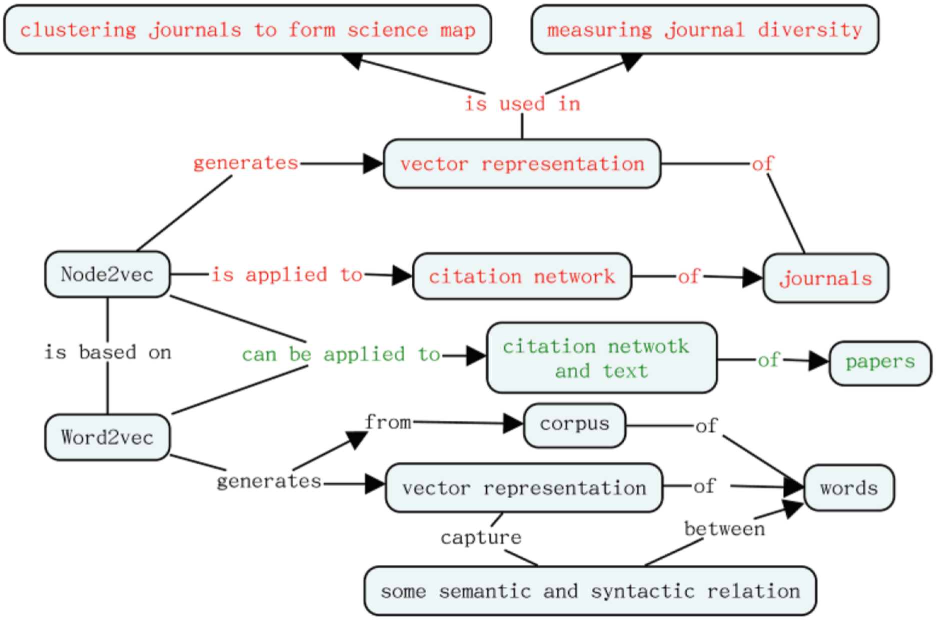

Figure 5. A graphic summary of our work: concepts and connections in red are the ones that have been implemented in the current work while the ones in green can be topics of future investigation. The rest of concepts and connections have been proposed and implemented in earlier studies, see for example, (Mikolov et al., 2013) and (Grover & Leskovec, 2016).

Acknowledgements

This work was supported by the NSFC under Grant No. 61374175, the China Postdoctoral Science Foundation under Grant 2017 M620944, and Fundamental Research Funds for the Central Universities.

Author contributions

Jinshan Wu (jinshanw@bnu.edu.cn), Liying Yang (yangly@mail.las.ac.cn) and Zhesi Shen (shenzhs@mail.las.ac.cn) designed this study. Shen and Fuyou Chen (chenfuyong@mail.las.ac.cn) performed the analysis. All participated in writing up the manuscript.

References

Boyack, K., Glänzel, W., Gläser, J., Havemann, F., Scharnhorst, A., Thijs, B., van Eck, N. J., Velden, T., & Waltmann, L. (2017). Topic identification challenge. Scientometrics, 111, 1223–1224.

Boyack, K. W., & Klavans, R. (2014). Including cited non-source items in a large-scale map of science: What difference does it make? Journal of Informetrics, 8, 569–580. doi:10.1016/j.joi.2014.04.001.

Colavizza, G., Boyack, K. W., van Eck, N. J., & Waltman, L. (2018). The closer the better: Similarity of publication pairs at different cocitation levels. Journal of the Association for Information Science and Technology, 69, 600–609. doi:10.1002/asi.23981.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics, 56, 357–367.

Glänzel, W., & Thijs, B. (2011). Using core documents for the representation of clusters and topics. Scientometrics, 88, 297–309.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855–864). ACM.

Haunschild, R., Schier, H., Marx, W., & Bornman, L. (2018). Algorithmically generated subject categories based on citation relations: An empirical micro study using papers on overall water splitting. Journal of Informetrics, 12, 436–447. doi:10.1016/j.joi.2018.03.004.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. Scientometrics, 75, 607–631.

JCR2017 (2018). 2017 journal impact factor, journal citation reports (clarivate analytics, 2018).

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. Journal of the American Society for Information Science and Technology, 60, 455–476.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? Journal of the Association for Information Science and Technology, 68, 984–998.

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the journal citation reports? Journal of the American Society for Information Science and Technology, 57, 601–613.

Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2017). Generating clustered journal maps: An automated system for hierarchical classification. Scientometrics, 110, 1601–1614.

Leydesdorff, L., Bornmann, L., & Wagner, C. S. (2017). Generating clustered journal maps: an automated system for hierarchical classification. Scientometrics, 110, 1601–1614. doi:10.1007/s11192-016-2226-5.

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2018). Betweenness and diversity in journal citation networks as measures of interdisciplinarityâ€"a tribute to eugene garfield. Scientometrics, 114, 567–592.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. Journal of Machine Learning Research, 9, 2579–2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In advances in neural information processing systems (pp. 3111–3119).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Rao, C. R. (1982). Diversity: its measurement, decomposition apportionment and analysis. Sankhy : The Indian Journal of Statistics, Series A, 44, 1–22.

Schakel, A. M., & Wilson, B. J. (2015). Measuring word significance using distributed representations of words. arXiv preprint arXiv:1508.02297.

Shen, Z., Yang, L., Pei, J., Li, M., Wu, C., Bao, J., Wei, T., Di, Z., Rousseau, R., & Wu, J. (2016). Interrelations among scientific fields and their relative influences revealed by an input—output analysis. Journal of Informetrics, 10, 82–97. doi:https://doi.org/10.1016/j.joi.2015.11.002.

Sjogarde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. Journal of Informetrics, 12, 133–152. doi:10.1016/j.joi.2017.12.006.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. Journal of the Royal Society Interface, 4, 707–719.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, 11, 2837–2854.

Waltman, L. (2016). A review of the literature on citation impact indicators. Journal of Informetrics, 10, 365 – 391. doi:https://doi.org/10.1016/j.joi.2016.02.007.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. Journal of the American Society for Information Science and Technology, 63, 2378–2392.

Zhang, L., Rousseau, R., & Glanzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. Journal of The Association for Information Science and Technology, 67, 1257–1265. doi:10.1002/asi.23487.